arXiv:2303.06078v1 [eess.AS] 10 Mar 2023

端到端　　图像到音频转换神经网络

刘晨[1,2], *Michael Deisher*[2], *Munir Georges*[3,4]

[1]俄勒冈健康与科学大学，美国俄勒冈州[2]英特尔公司，
美国俄勒冈州希尔斯伯勒[3]英特尔实验室，德国慕尼黑[4]
因戈尔施塔特应用技术大学 chliu@ohsu.edu,{
michael.deisher,munir.georges}@intel.com

摘要

本文介绍了一种端到端（E2E）神经网络架构，用于在资源受限的个人计算设备上渲染少量显示内容的音频。该架构旨在从硬件层面解决视力受损或视觉分心用户的无障碍问题。本文回顾了神经图像到文本识别（ITT）和文本转语音（TTS）方法，并介绍了一种新技术，可高效地将两者整合，兼具效率和反向传播能力，从而构建出高效且可训练的非自回归端到端图像转语音（ITS）神经网络。实验结果表明，与非端到端方法相比，所提出的端到端系统速度提高了 29%，参数减少了 19%，但音素准确率降低了 2%。文中还提出了提高准确率的未来研究方向。

索引词— 光学字符识别、文本到语音、图像转语音

1. 引言

触摸屏个人计算设备的用户，例如手机、平板电脑和笔记本电脑，可能会遇到因安全考虑或视力障碍而难以获取显示内容的情况。操作系统（OS）的无障碍功能可以朗读屏幕上的文本，从而缓解这一问题。然而，屏幕阅读器通常从属于操作系统，可能无法呈现图像中的文本内容。使用专用神经网络协处理器实现此类功能，具有每瓦成本低、功耗低、能够抵御操作系统故障且不依赖应用程序等优点。虽然可以利用独立的图像识别和音频生成神经网络，从显示区域生成音频，但非自回归端到端（E2E）神经网络架构更适合此应用，因为它能简化硬件设计和推理过程。为了降低功耗和成本，还必须尽量减少所需的内存占用。在

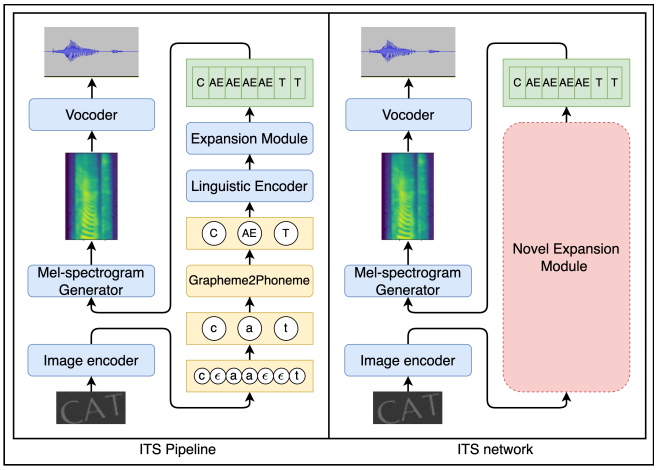

图 1。图像到音频：黄色标示的不可反向传播的后处理／预处理步骤，由图中红色所示的新型扩展模块取代。由此可实现端到端训练。

本文介绍了一种非自回归端到端神经网络，适用于在个人计算设备的嵌入式硬件上实现图像转语音子系统。我们基于图像到文本识别（ITT）和文本转语音（TTS）领域的既有研究，提出了一种新方法，可将二者整合为一个可训练的图像到语音神经网络。据我们所知，这是首次实现这一点。

图1所示的网络结构根据给定的固定尺寸单词图像合成语音。图的左半部分展示了采用独立的非自回归图像到文本和文本到语音模型的流水线（详见第2节）。我们希望构建一个可完全进行反向传播的非自回归端到端网络。然而，图中以黄色标出的后处理／预处理步骤无法进行反向传播。因此，我们将工作重点放在图1右半部分的粉色方框上，该部分处理语音与编码后图像之间的对齐问题。我们的贡献是为端到端图像到语音系统提出一种直观的序列扩展模块，它具有足够的灵活性，能够适配多种图像编码器和梅尔频谱图生成器。

本文其余部分安排如下：第2节结合以往研究介绍图像到文本识别问题。第3节详细介绍端到端深度神经网络架构。第4节介绍训练过程和评估指标。最后，第5节给出实验结果并总结全文。

2. 背景

图像到文本识别（ITT）旨在识别输入图像中的文本。通常，ITT 包含三个模块：可选的矫正器、图像编码器和序列解码器。矫正器通过以统一的方式转换各种类型的文本不规则性，对图像进行分割和归一化 [1]。图像编码器从归一化后的图像中提取隐藏表示 [1]。解码器则根据隐藏表示生成字符序列。非自回归图像到文本可预测任意长度的序列。在训练过程中，真实文本会通过添加额外的占位符（即 ϵ 或重复字符）进行扩展，使扩展后的真实文本长度与模型的隐藏表示序列相同，然后再将其输入所选的损失函数 [2, 3, 4]。推理时，移除占位符后即可得到字符。不同损失函数采用的扩展算法各不相同。连接主义时序分类（CTC） [2] 同时使用重复字符和 ϵ。聚合交叉熵 [5] 仅使用 ϵ 作为占位符。这两种损失函数都会最小化目标文本与所有可能的扩展文本之间的距离，因此，占位符可以出现在原始输出的任意位置。Cai 等人 [6] 改进了一种传统方法：在真实文本末尾插入 ϵ 以扩展文本，并通过交叉熵（CE）损失最小化该扩展文本。Cai 等人 [6] 指出，经过适当配置后，这种方法能够达到与 CTC 相当的性能 [2]。因此，我们在此采用这种方法。

文本转语音（TTS）旨在根据文本合成自然且清晰易懂的语音。它首先将文本（即字素或音素序列）转换为声学特征（例如梅尔频谱序列），然后通过声码器将声学特征转换为音频样本。由于音素序列远短于其对应的梅尔频谱序列，因此音素与梅尔帧之间的对齐至关重要。近期解决这一问题的方法包括使用自回归深度神经网络，以及利用音素到梅尔帧对齐的单调性、显式预测时长以弥合长度差异的非自回归深度神经网络 [7]。这些序列扩展模块包含扩展长度调节器和音素时长预测器，用于实现音素与其对应梅尔帧之间的硬对齐。真实对齐由外部对齐器获取。最早提出的非自回归模型 FastSpeech [7] 和 ParaNet [8], 利用预训练自回归模型获取音素级真实对齐，而 FastSpeech2 [9] 则利用强制对齐器实现相同目的。PortaSpeech [10]

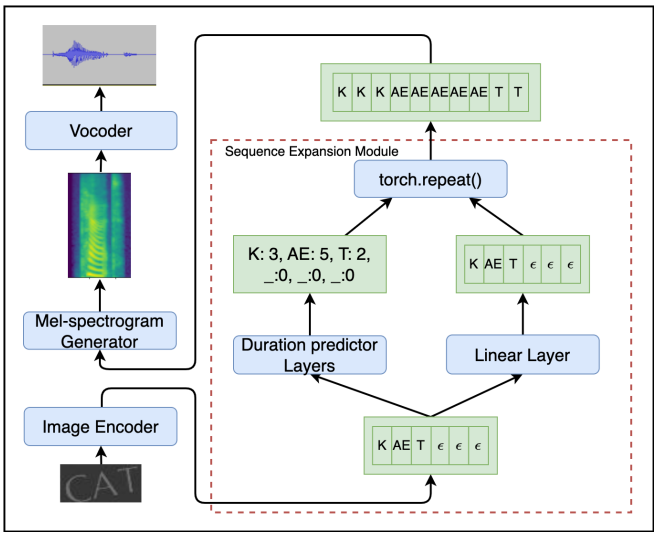

图 2。处理一张包含单词“CAT”的图像。符号 ϵ 表示占位符。绿色矩形框表示符号的隐藏表示。

通过从外部对齐器学习词级硬对齐，并通过注意力机制学习音素级软对齐，减轻了音素对齐不精确带来的负面影响。Parallel Tacotron2 [11] 和 ESA [12] 结合使用可微时长预测器与基于注意力的长度调节器，对话语时长进行建模，并在不使用外部对齐器的情况下学习音素对齐。Glow-TTS [13] 利用基于流的模型的可逆变换特性，通过动态规划搜索最可能的单调对齐。在本研究中，我们采用以音素序列为输入的非自回归文本转语音。由于 [12] 中报告的问题，我们避免使用字素序列作为输入。

3. 端到端图像转语音架构

端到端系统的核心挑战是将梅尔频谱与任意长度的编码图像序列对齐。在文本到语音（TTS）中，每个梅尔频谱序列的长度直接取决于编码后的音素序列。然而，在图像到语音（ITS）中，编码图像序列的长度与梅尔频谱序列的长度并无关联。为弥合这一差距，我们引入一个非音素符号 ϵ，作为唯一的占位符，并在音素序列末尾插入 ϵ，将其扩展至任意固定长度。此外，我们规定，只有音素可以与梅尔频谱对齐，任何 ϵ 的时长都必须为零。总之，我们为时长预测器分配两项任务：1）预测音素时长（正值）；2）识别 ϵ（零值）。这样便无需为第二项任务添加额外层（例如 [14] 中使用的层）。

我们以图 2 中的一张示例图像来展示我们的架构。

首先，图像编码器将“cat”编码为扩展后音素序列的隐藏表示。然后，时长预测器预测序列中每个向量的时长。如果隐藏向量表示的是 ϵ，预测器就会生成零。此外，右侧的线性层会将隐藏表示转换为梅尔频谱生成器所需的维度。第三，我们通过将每个向量重复 d 次来扩展转换后的表示，其中 d 是其对应的时长。第四，梅尔频谱生成器以扩展后的表示为输入，并合成梅尔频谱图。最后，声码器根据合成的梅尔频谱图合成原始波形。

图像编码器。我们沿用第2节介绍的 ITT 架构。我们使用与 Shi 等人相同的矫正器 [1]，并采用 HBONet [15] 作为骨干网络，从矫正后的图像中提取隐藏特征。池化层从隐藏特征中提取全局语义信息，并将其输入 26 个线性层。i^{th} 线性层用于预测i^{th} 输出。如果一个单词只有 N 个音素，那么最后 26-N 个层应预测 ϵ。我们的 HBONet 配置与 [15] 相似，只是将第一个倒置残差块的 n 和 s 的值改为 1（见 [15]中的表 1）。经验表明，这项修改对模型准确率的影响很小，同时可减少 0.5M 个参数。

时长预测器。时长预测器由两个卷积块组成。每个卷积块包括一个 1D 时域-通道可分离卷积、一个 1×1 逐步卷积、层归一化、ReLU 和随机失活。线性层与 softplus 层将隐藏表示序列映射为标量序列，这些标量即为预测的音素时长。

梅尔频谱图生成器。为了采用轻量级架构，我们使用与 PortaSpeech [10]提出的相同的基于变分自编码器（VAE）的合成器。该合成器由一个编码器、一个保体积（VP）流先验模型和一个解码器组成。编码器由一个步幅为 4 的一维卷积构成，后接 ReLU 激活、层归一化和非因果 WaveNet [16]。先验模型是一个保体积归一化流，由一个残差耦合层和一个按通道翻转操作组成。解码器由一个非因果 WaveNet [16] 和一个步幅为 4 的一维转置卷积构成，后接 ReLU 和层归一化。

4. 训练与评估

系统采用多步方式进行训练，具体如下。为了训练图像编码器，我们使用交叉熵损失，确保编码器将图像转换为音素的隐藏表示和 ϵ，并采用与 Cai 等人相同的训练配置 [6]。接下来，我们冻结预训练的图像编码器，并按照与 PortaSpeech 相同的配置训练扩展模块和梅尔频谱图生成器 [10]。最后，我们使用原始波形和真实梅尔频谱图训练声码器。因此，声码器的训练是独立的。我们为 HiFiGANv2 采用 HiFiGAN [17] 的配置，其模型仅包含 0.9M 个参数。

4.1. 训练数据集与预处理

输入图像的尺寸设为 64x224。由于我们使用为 ITT 任务设计且仅提供图像—单词对的数据集，因此采用字素到音素（G2P）工具 [18] 将单词转换为音素序列。为确保结果可复现，我们使用 Microsoft®Azure 语音服务合成真实音频。由于采用多步训练过程，我们将介绍每个步骤所用的训练数据。

图像编码器。我们将两个数据集合并：MJSynth (MJ) [3], , 其中包含从一个含有 9 万个英语单词的词库生成的 900 万张单词框图像，以及 Synthtext[19]。我们排除包含非字母字符的图像。

序列扩展模块与梅尔频谱图生成器。我们为 MJ 中的每个单词生成真实音频。我们从清理后的 MJ [3]中抽取每个单词对应图像的 20%。此外，为了增加短单词的样本量，我们为每个字符数少于 6 的单词从 Synthtext [19] 中抽取 100 张图像。为了增加每个单词的速度变化，我们对每个图像-音频对随机应用速度扰动。根据经验，这一操作能够减轻外部对齐器造成的对齐错误的影响。我们将采样率为 22050 Hz 的原始波形转换为梅尔频谱图，帧大小为 1024，帧移为 256。该数据集同时包含图像和单词信息，因此我们可以训练图像到语音和文本到语音模型。

声码器。我们为 LJSpeech [20] 中的每个话语生成真实音频，并将合成的音频用作训练数据。

4.2. 测试数据集

为了让端到端图像到语音系统适用于个人计算设备上可能显示的内容，我们使用开源工具包 [1]，根据维基百科中的单词频率合成出现频率最高的 3000 个单词，并采用随机字体和配色组合。

4.3. 评估指标

根据实证结果，我们发现 Microsoft®Azure 自动语音识别（ASR）服务在真实音频录音上的表现良好。因此，我们使用该服务转写合成的输出音频。然而，由于每段合成输出仅包含一个单词，而且自动语音识别无法在缺少上下文的情况下区分同音异义词，我们的

[1]github.com/Belval/TextRecognitionDataGenerator

评估基于音素序列，而非字符序列。我们采用两项评估指标：音素错误率（PER）和词准确率。

5. 实验

5.1. 语音合成质量

我们比较了端到端图像文本语音合成模型与非端到端图像到语音流程的性能。端到端架构的一个潜在优势是可以减少参数数量。在标准的非端到端图像到语音流程中，有两个编码器：一个用于文本识别的图像编码器、一个用于梅尔频谱图合成的语言编码器，以及一个字素到音素转换器。我们的端到端模型使用单个图像编码器对图像进行编码，以合成梅尔频谱图。

我们构建了一个非端到端图像到语音流程作为基线。我们采用图像到语音中的图像编码器所使用的相同架构，使用 MJSynth [3] 和 SynthText [19]训练 ITT 模型。基于变分自编码器的文本转语音 [21] 使用与端到端图像文本语音合成模型相同的数据集进行训练。表 1 展示了 PER、准确率和参数规模。文本到语音系统在底层采用嵌入层，将输入的音素索引转换为音素嵌入。音素的嵌入具有通用性。另一方面，我们的图像到语音系统使用图像编码器的隐藏表示来表示音素，而这种表示会受到输入图像相关变化的影响。基于这一差异，我们认为音素表示的鲁棒性不如嵌入。未来的研究之一将评估表示的鲁棒性如何影响准确率。

5.2. 推理速度

由于我们采用非自回归模型以实现快速推理，因此将本模型的速度与第 5.1 节所用的非端到端流水线进行比较。我们使用识别速度（每秒图像数）和实时率（RTF）来评估推理速度。前者用于 ITT 领域，后者用于文本到语音领域。由于图像到语音流水线包含两个深度神经网络，我们将给定输入在两个深度神经网络上的推理时间之和视为流水线的推理时间。我们使用 Nvidia 2080Ti，以批量大小为 1 合成梅尔频谱图。表 2 显示，端到端图像到语音的速度快于图像到语音流水线。

5.3. 数据分布的影响

如今的文本到语音研究主要关注句子级性能 [7, 9, 10, 12, 11]以及长语句的鲁棒性 [22]，很少研究孤立词的清晰度。然而，由于我们希望将图像到语音系统用于朗读孤立词，清晰的发音至关重要。我们注意到，图像到语音和基于变分自编码器的文本转语音模型在合成包含更多

模型名称	PER (%)	Acc (%)	参数
端到端图像到语音	4.7	87.8	6.1M
非端到端图像到语音流程	2.7	92.3	7.5M

表 1。图像到语音系统的音素错误率和词准确率。不包括声码器。

模型名称	速度（图像/秒）	RTF
端到端图像到语音	78.0	0.0167
非端到端图像到语音流程	59.9	0.0236

表 2。推理速度对比。

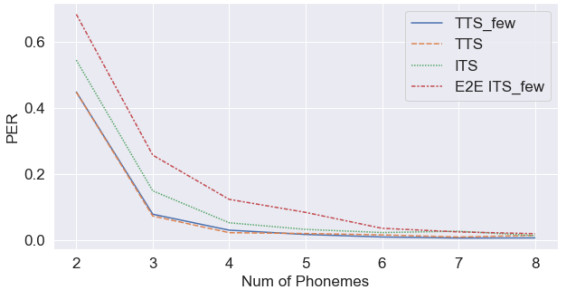

图 3。我们根据音素长度对所有测试样本进行分类，并展示模型在音素长度为 2 至 8 时的性能。

音素的单词时表现更好。若从 MJSynth [3], 为每个单词抽取数量相近的图像，则 6+音素词的总样本量会更大。我们认为，音素长度的样本分布是影响模型性能的一个关键因素。

为了定量评估这一影响，我们使用包含较少短单词样本的训练集训练了新的 ITS 和 TTS 模型，并分别将这些模型称为 *E2EITS few* 和*TTS few*。我们移除了所有来自 Synthtext [19]的短单词样本。图 3 显示，与对应模型相比，使用更多短单词样本训练的 E2E ITS 性能明显更好，而较长单词上的性能则略受影响。此外，与 TTS 相比，ITS 从额外的短样本训练中获益更多。

6. 结论

我们通过引入一种直观的序列扩展模块，提出了首个非自回归端到端图像到语音系统。该模块足够灵活，可适配多种图像编码器和梅尔频谱图生成器。我们的图像到语音模型足够小巧，可部署在移动设备上。此外，我们展示了数据分布对于训练图像到语音模型的重要性。未来，我们将增强图像到音素表示的通用性。

7. 参考文献

[1] Baoguang Shi、Xinggang Wang、Pengyuan Lyu、Cong Yao 和 Xiang Bai，“通过自动校正实现鲁棒的场景文本识别”，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第4168–4176页。

[2] Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber，《联结主义时序分类：使用循环神经网络标注未分段序列数据》，载于第*23*届国际机器学习会议论文集，2006年，第369–376页。

[3] Max Jaderberg、Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman，《用于自然场景文本识别的合成数据与人工神经网络》，载于*NIPS* 深度学习研讨会，2014年。[4] Max Jaderberg、Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman，《使用卷积神经网络识别自然场景中的文本》，《国际计算机视觉杂志》，第116卷，第1期，第1–20页，2016年1月。[5] Zecheng Xie、Yaoxiong Huang、Yuanzhi Zhu、Lianwen Jin、Yuliang Liu 和 Lele Xie，《用于序列识别的聚合交叉熵》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2019年，第6538–6547页。[6] Hongxiang Cai、Jun Sun 和 Yichao Xiong，《重新审视场景文本识别的分类视角》，*arXiv*预印本 *arXiv:2102.10884*，2021年。[7] Yi Ren、Yangjun Ruan、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech：快速、稳健且可控的文本转语音》，《神经信息处理系统进展》，第32卷，2019年。[8] Kainan Peng、Wei Ping、Zhao Song 和 Kexin Zhao，《非自回归神经文本转语音》，载于国际机器学习会议。PMLR，2020年，第7586–7598页。[9] Yi Ren、Chenxu Hu、Xu Tan、Tao Qin、Sheng Zhao、Zhou Zhao 和 Tie-Yan Liu，《FastSpeech 2：快速、高质量的端到端文本转语音》，*arXiv*预印本 *arXiv:2006.04558*，2020年。

[10] Yi Ren、Jinglin Liu 和 Zhou Zhao，《Portaspeech：便携式高质量生成式文本转语音》，《神经信息处理系统进展》，第34卷，第13963–13974页，2021年。

[11]Isaac Elias、Heiga Zen、Jonathan Shen、Yu Zhang、Ye Jia、RJ Skerry-Ryan 和 Yonghui Wu，“Parallel Tacotron 2：一种采用可微时长建模的非自回归神经文本转语音模型”，*arXiv* 预印本 *arXiv:2103.14574*，2021年。

[12] Jeff Donahue、Sander Dieleman、Mikołaj Bi´nkowski、Erich Elsen 和 Karen Simonyan，“端到端对抗式文本转语音”，*arXiv* 预印本 *arXiv:2006.03575*，2020年。[13] Jaehyeon Kim、Sungwon Kim、Jungil Kong 和 Sungroh Yoon，“Glow-TTS：通过单调对齐搜索实现文本转语音的生成流”，发表于《神经信息处理系统进展》，第33卷，第8067–8077页，2020年。[14]Isaac Elias、Heiga Zen、Jonathan Shen、Yu Zhang、Ye Jia、Ron J Weiss 和 Yonghui Wu，“Parallel Tacotron：非自回归且可控的文本转语音”，载于*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2021年，第5709–5713页。[15] Duo Li、Aojun Zhou 和 Anbang Yao，“HBONet：两个正交维度上的和谐瓶颈”，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第3316–3325页。[16] Aaron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior 和 Koray Kavukcuoglu，“WaveNet：一种用于原始音频的生成模型”，*arXiv* 预印本 *arXiv:1609.03499*，2016年。[17] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae，“HiFi-GAN：用于高效、高保真语音合成的生成对抗网络”，发表于《神经信息处理系统进展》，第33卷，第17022–17033页，2020年。[18] Kyubyong Park 和 Jongseok Kim，“g2pe”，https: //github.com/Kyubyong/g2p，2019年。[19] Ankush Gupta、Andrea Vedaldi 和 Andrew Zisserman，“用于自然图像文本定位的合成数据”，载于*IEEE*计算机视觉与模式识别会议，2016年。[20] Keith Ito 和 Linda Johnson，“LJ语音数据集”，https://keithito.com/ LJ-Speech-Dataset/，2017年。[21]Liu Yi，“NATSpeech”，https://github.com/NATSpeech/NATSpeech，2022年。[22] Rohan Badlani、Adrian La´ncucki、Kevin J Shih、Rafael Valle、Wei Ping 和 Bryan Catanzaro，“一种对齐统领所有文本转语音的方法”，载于*ICASSP 2022-2022 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2022年，第6092–6096页。